Paper

The Integrated Data Environment: a new tool for interoperability and effective data integration for command and control

Jon Wilkes

Abstract — This paper describes an ongoing effort at NC3A to provide one integrated database which contains data from a number of different sources. Initially, these sources are legacy NATO systems. Later, other systems, including messaging interfaces of a wide variety, and national systems, will be added. A common data model is used as the lingua franca between systems. A COTS product has been identified that creates translator boxes to provide interfaces to and from the legacy systems.

Keywords — Integrated Data Environment, data interoperability, common data model, data modelling, data translation, ATCCIS, LC2IEDM, ADatP-32.

1. Introduction

The NATO C3 Agency has responded to customer requirements with the Integrated Data Environment (IDE) project, which has been evolving over the past three years. The intention of the effort is to provide one integrated database which contains data from a number of different sources; in the first place these will be legacy internal NATO systems. Later, other systems, including messaging interfaces of a wide variety, and national systems, will be added as requirements and political concurrence allow. It is foreseen that IDE will play a significant role in the core capability package for the Bi-SC AIS. This paper is based heavily on the work started by the late Martin Krick.

2. The problem

Many of the data exchange problems that have confronted and bedevilled NATO for the past few decades have arisen from the fact that early systems were conceived, developed and implemented as stand-alone, or "stovepipe", systems by groups of users and technicians whose requirements horizon extended no further than the immediate needs of the system on which they were engaged. In the early days, interoperability of data models was not even considered relevant.

As time progressed, and the initial desirability of being able to pass information from one system to another became a more firm requirement, many mechanisms were devised to address these issues, but always with the caveat that the software within the in-service systems, seen to be of such acquisition cost as to be untouchable for interoperability needs, could not be modified to assist in the process of bringing systems together to provide for any meaningful direct exchange of data. In addition, because early systems were so expensive, and therefore made available only to the smallest possible community of users, and because many of the more senior users had no ADP facilities at all, or at most a simple teletype, these early mechanisms were specified to be able to be used in manual environments, leading to the definition of a range of messages. Once again, these message definitions were aimed at encapsulating the specific needs of the group of users responsible for the definition of each message; correlation between messages was not a driving force in the definitions.

3. Previous studies

Many studies were carried out when the nature of the problem became so large that it could no longer be ignored; these studies stressed the need for common standards for data definition, but could not provide low-cost solutions and their conclusions were therefore ignored. In essence, they proposed a "data fusion" approach, which is nowadays seen to be both impractical and unnecessary.

4. The data fusion approach

The principle behind a data fusion approach is to define a single data model, and implement a single database, which will encompass the entire set of data currently held in all existing systems. This approach has some advantages, but also has many more major drawbacks which make it an impractical proposition. If we take two or three existing systems, and create a new database which holds all the data previously held in the three individual databases in accordance with a new all-encompassing data model, then the new database will not be the same as any of the old ones. Each application suite in the original systems must therefore be re-written.

It might be possible to create a database interface package for each system to make the new database appear as the old database, but that too would be substantial effort (and there would be no *ab initio* guarantee of feasibility) and would represent an additional load for the original system which it might well be ill-suited to handle.

A further major, and potentially even more serious, disadvantage is that if another legacy system were to be added to the fusion set, it may impose changes on the data model which would have a knock-on effect on all current systems within the fusion set. This would lead to potentially exponentially soaring costs, and to management problems of equally soaring complexity. Little wonder that the NATO committees of the time were not persuaded to follow down this route!

The perceived advantages and disadvantages of the data fusion approach can be summarized as:

- single view of all data,
- single physical database from which all applications can draw data,

whereas the disadvantages are:

- need to agree the (large) data model between 19 nations and all NATO HQs and Agencies,
- immediate impact on all legacy systems which are required to conform to the new global data model,
- applicable only for a small number of systems (three or maybe four),
- ongoing management overhead for the fusion schema,
- complexity increases dramatically with the number of systems,
- process becomes unmanageable with large numbers of systems.

It should be noted that the advantages are not matched by any known requirement for all data to be perceived in a single view, nor that there should be a capability of providing a single database implementation which would hold all data; these advantages represent theoretical technical possibilities only. By contrast, the listed disadvantages are very real, not least the political problems associated with the first of those listed. Corresponding agreements in related areas are not famous for the speed with which such agreements have been reached nor for the technical clarity of the final agreements.

5. Other more recent studies

In the last ten years, other initiatives have been taking place on a lower profile basis, and the fruits of their endeavours are now beginning to become visible in a number of places; national implementations based on these initiatives have been put in place and have become sufficiently mature for reasonable projections to be made. Principal of these initiatives is the multi-nation ATCCIS[1] study, sponsored and led by NATO, with active participation at varying levels by eleven nations.

The major outputs of the ATCCIS study to date have been:

- a wealth of well-documented analysis,
- a fully specified data model for information exchange,
- an ATCCIS replication mechanism (ARM) for selective transfers of data between two or more ATCCIS-conformant databases.

The primary achievement of the data modellers is that they recognised that they were endeavouring to specify a data model to facilitate the exchange of information rather than for the design or development of systems; thus the level of detail of the model is appropriate to information exchange, and much low-level data, which would typically be found only in specialist systems, was not included. This separation of "local" data and "global" data has been one of the foundation links of the NC3A work on the Integrated Data Environment.

6. The integration approach

The separation of local and global data leads immediately to the concept of an IDE which addresses only some of the totality of data held in all existing (and future) systems. It also leads directly to the recognition that the IDE can be established (either as a virtual database or as a real one) for new purposes, and that the existing systems can be left with their current databases and database management systems – be they rudimentary or advanced – with the immediate benefit that no changes to those systems are required. Indeed, it became one of the design objectives of the IDE work that the IDE concept should be seen to be non-intrusive from the perspective of any legacy system.

In the integration approach, data are translated from the native (legacy) environment to the common data model of the IDE, so that the translated data subsets reside in a single database or transmission mechanism with one common data model describing all data. We may think of this common data model as a "lingua franca".

The integration approach offers as advantages:

- single view of all global data,
- no impact on legacy systems,
- no requirement to have a single database,
- all future applications can draw global data from existing databases,

[1]The common ATCCIS generic hub 4 data model was forwarded to NATO in 1999. NATO initiated a standardisation process for this data model, now called the Land C2 Information Exchange Data Model (LC2IEDM). The respective STANAG 5532 (ADatP-32) has been submitted as draft and is expected to be agreed in 2001.

- process remains manageable with large numbers of systems,
- ongoing management overhead for the integrated database is much smaller than for the fusion approach,
- technology is mature and in use in large commercial organizations,

and as disadvantage:

- as of end 2001, the technology has not been proven within a NATO operational system (but a demonstrator has been produced, and is clearly scalable to full operational use).

It may be seen that almost all of the disadvantages of the fusion approach have been stood on their heads for the IDE approach. The single view of all data, which was never supported as an operational requirement, has been scaled down to become a single view of all global data, for which operational requirements most certainly exist. The previous high impact, in terms of both cost and operational implications, of the fusion approach, has become a zero impact on those systems. The management problems remain tractable.

On the disadvantage side, the technology has not yet been tested in a full NATO operational environment, but a four-system demonstrator has been produced, and the technology is scalable to encompass a very large number of systems. In particular, the technology ensures that the management problems remain at the one-system level, and therefore do not grow as the number of systems being integrated expands.

7. Alternative techniques

There are two techniques available to implement the IDE function, data mediation and data translation. Data mediation works by first making associations of the meta-data of the data sources and the data sink, and then automatically converting source data to the sink on the basis of these pre-determined associations. In principle, this is a very powerful technique; however, at the present time the technology is still in the research stage, with academic institutions producing small-scale demonstrations. No proposals for a full-scale demonstration have come to our notice at this time. The technology is thus considered to be far too immature to be considered for introduction to NATO at the present.

By contrast, data translation is a very much more mature technology which has been in use in commerce for some time. Most of those applications have been for data warehousing applications, but some applications have been for genuine data integration applications. Where the translation process is carried out on a one-translator-per-system basis, there are very few problems about scaling to multiple systems. The scaling problems are mainly associated with the suitability of the sink data model for the spread of data types to be found in the source systems; in this respect, the highly generic nature of the ATCCIS data model is of immense benefit in minimising such risks. Finally, it must be emphasised that both techniques act on the conversion of data on a one-for-one basis. Data aggregation, data fusion and other application-level functions are outside the scope of both technologies.

8. The IDE architecture

Figure 1 gives a very simplified overview of the IDE architecture resulting in the use of translation techniques on a translator-per-system basis. Data from each legacy or national system is processed by its own local translation process to the target (sink) data model and added to the data model of the target system by normal database update techniques. The translation mechanism is a process, implemented as a software package; although for simplicity it is shown in this slide as though it were a separate system, it could equally well be hosted on the legacy system if that were to prove to be the preferred option. However, to emphasise the "No impact" concept, we always show it as a separate system.

Because the translator process will only translate data about which it has been provided with appropriate translation data (which is another form of meta-data), it acts as a simple form of guard against the accidental translation of data which is not to be released. However, the translator process makes no claims to be an approved guard, and additional security devices would normally be expected to be fitted by national authorities to protect national systems which may contain nationally-sensitive data. These would typically be positioned between the national system and the translator.

Both the initial configuration of the translator, and any subsequent upgrades or changes to a national system will require detailed analysis of the source system in order to specify the translation meta-data. For this reason, the configuration of the translators is expected always to be done by the nation concerned. Figure 1 thus shows the translators residing in the national management domain, with the exception of the specification of the output format (ATCCIS conformant) which is essentially public domain.

9. Work done by NC3A

The preliminary study on data mediation carried out in 1998 showed that the technique held potential for complex translation situations, and for the tracking of changes to databases. A simple demonstration system was created, using the most rudimentary meta-data, which was shown at JWID-99. Much interest was demonstrated by visitors at the ability to show data from three different systems out of a common database in response to a single query, with the consequential ability to provide for integrated data solutions.

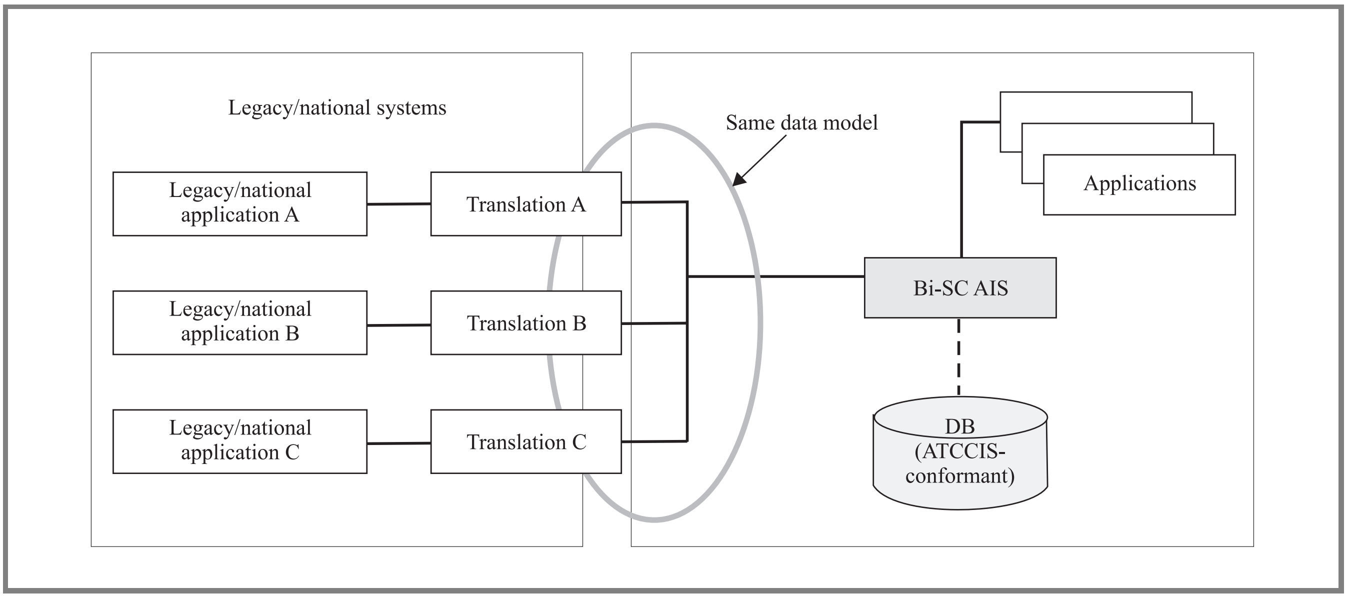


Fig. 1. The IDE architecture.

Evaluation of a contractor report made clear that, although the concepts behind the data mediation technology were both powerful and useful, the technology was very immature with no commercially available implementations of a data mediator product, and little prospect of any such products appearing in the market for some considerable time. Data mediation may have benefits for special situations in the future, yet to be assessed and proven.

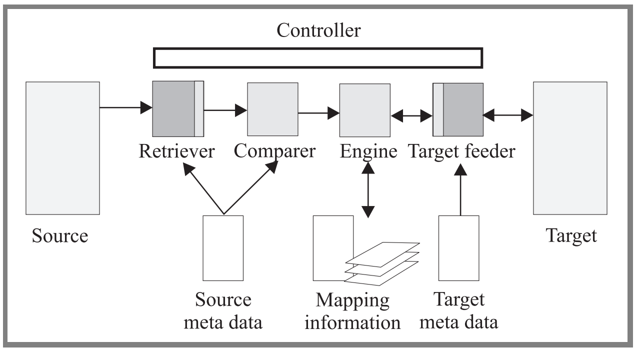


Fig. 2. The architecture of the translator box.

At the same time, an investigation was made of other products, all of which proved to be data translator systems, and it was determined that this offered a better approach for the near term. A contract was let for the development of a demonstrator using translation technology for display at JWID 2000. Problems with the suitability of the translation proposed by the contractor meant that only a very limited demonstration could be mounted at that time, but a very good tool has since been developed by the contractor as a COTS product, which has proven to be very successful and very flexible. A demonstration held at NC3A in late November 2000 showed the capabilities of this tool, and the design gives confidence for its use in many other situations, including message-oriented environments. A major demonstration was held at JWID 2001 and JWID 2002 at SHAPE.

Figure 2 shows the architecture of the translator box produced by the tool.

10. The selected data model for IDE

The selection of the ATCCIS data model, in the form known as the SHAPE Land Command and Control Information Exchange Data Model (LC2IE DM) proved to be a sound choice. The complex nature of this data model means that the specifications of the translations are themselves more complex, but no instances were found in the work on the four NATO legacy systems where translations could not be specified with alacrity and accuracy.

The NATO data administration group reference model is also based on the ATCCIS model, and is under strict configuration management; the LC2IE DM should similarly be placed under CM while it is being used as an interim measure before the full availability of the NATO reference data model. At the same time, some of the work of the NDAG could usefully be retro-fitted to the LC2IE DM to make it into a joint product, a JC2IE DM; the experience of NC3A and their contractor suggests that the minimal changes for the interim product would be small and easy to define and implement. For the November 2000 demonstration mentioned above it was necessary to add only four low-level entities (naval unit, air unit, naval facility and air facility) and to extend the range of a set of domain values to cover maritime and air factors. The total work took less than a couple of days; to repeat this work under full CM control would take less than one week. The future ATCCIS generic hub 5 may address the problem.

11. The tools used for IDE development

Mention has already been made of the shortcomings of the original analysis tools proposed by the contractor. These tools were designed for data warehousing applications where the primary focus of the tools was to analyse data – often dirty data – for which a data model did not exist. In the IDE situation, data models existed and were well documented (although there were some instances where the semantics of the data were not fully defined). Additionally, in data warehousing applications, the emphasis on fitting all source data into a single data model in the destination system does not apply. It is thus not surprising, with the benefit of hindsight, that the tools were found to be unsatisfactory for the IDE situation.

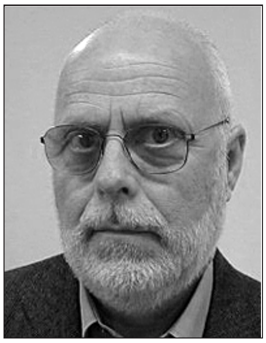
The analytical process involved in determining the translations required is both a skilled process and one which requires time. An analyst familiar with both the source system and the destination data model can complete several source tables each day if the source data model is "clean" and the semantics are fully defined and supported by exemplar data samples. Loose source data models, or a lack of semantic definition, or a lack of sample data, will slow the process to a considerable extent. The tool developed by the contractor provides considerable assistance in converting the results of the analysis into translation rules; future versions are expected to provide some additional assistance to the analysis itself, but cannot fully replace the need for analysis or the analyst.

12. Conclusions

NATO and the nations still have a plethora of incompatible data systems which are likely to remain in service for many years. A fusion approach is not appropriate, and is likely to prove unmanageable and unaffordable.

The Integrated Data Environment provides a response to this information management challenge that is both manageable and affordable, and is eminently suitable for an incremental growth approach.

Commercial off-the-shelf tools are available which support IDE and thus support coalition interoperability, NATO to NATO interoperability, NATO to nations interoperability, and coalition HQ to nations interoperability.



Jon Wilkes is a Senior Analyst Programmer in the Communications and Information Systems Division at the NATO C3 Agency in the Hague. He has been specialising in the interoperability of C2 systems, and the associated problems of the definition of data, for several years at the NC3A. More recently he has been assisting investigations into ways of implementing mechanisms for providing for interoperability between non-compatible systems which have led to the development of the IDE concept and the creation of tools to support the concept.

e-mail: jon.wilkes@nc3a.nato.int
NATO C3 Agency
Postbox 174
2501 CD The Hague, The Netherlands